Міністерство освіти і науки України Національний технічний університет України
"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ

Комп’ютерний практикум №1

«Експериментальна оцінка ентропії

на символ джерела відкритого тексту»

Виконав

студент групи ФБ-93

Флекевчук Данило

Перевірила

Селюх П.В.

Київ – 2021

Мета:

Засвоїти поняття ентропії та надлишковості мови. Навчитись проводити частотний аналіз тексту та за його результатами рахувати ентропію тексту. Покращити навички програмування.

Завдання:

- уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- написати програму для підрахунку частот літер і частот біграм в тексті, а також підрахунку H1, H2 за безпосереднім означенням. Підрахувати частоти літер та біграм, а також значення H1, H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1 Мб), де ймовірності заміняються відповідними частотами. Також отримати значення H1, H2 на тому ж самому тексті, в якому вилучено всі пробіли.
- за допомогою програми CoolPinkProgram оцінити значення H(10), H(20), H(30).
- використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Хід роботи:

I.

В цій частині роботи я рахую частоту входження літер, біграм(двох типів) у текстовий файл, та за цими показниками рахуємо ентропію мови. Для обчислення було використано мову програмування Python, так як вона найкраще підходить для важких обчислень. Для підрахунку входжень я використовував модуль collections а саме Counter для підрахунку входжень у текст та OrderdDict для сортування. Я розбив текст на різні типи біграм, з допомогою циклів з різними кроками (перехрестні: 2, прості: 1).

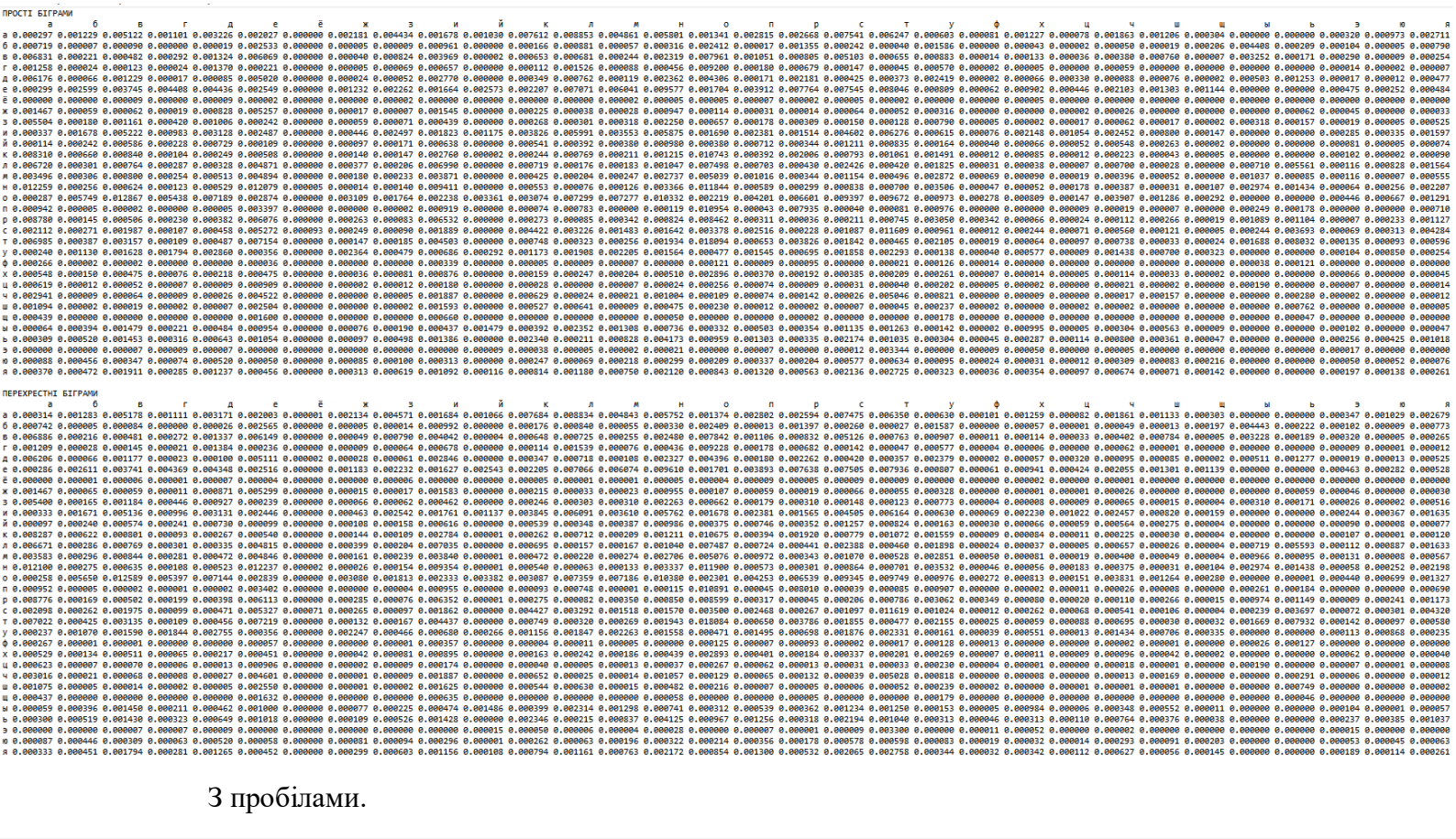
Код знаходится в тій же папці що і протокол.

Далі я порахував ентропію для біграм та літер, і посортував літери за зростанням частоти. І вивів розраховані значення ентропії в консоль. Ентропію я рахував за формулою, що була наведена в завданні.

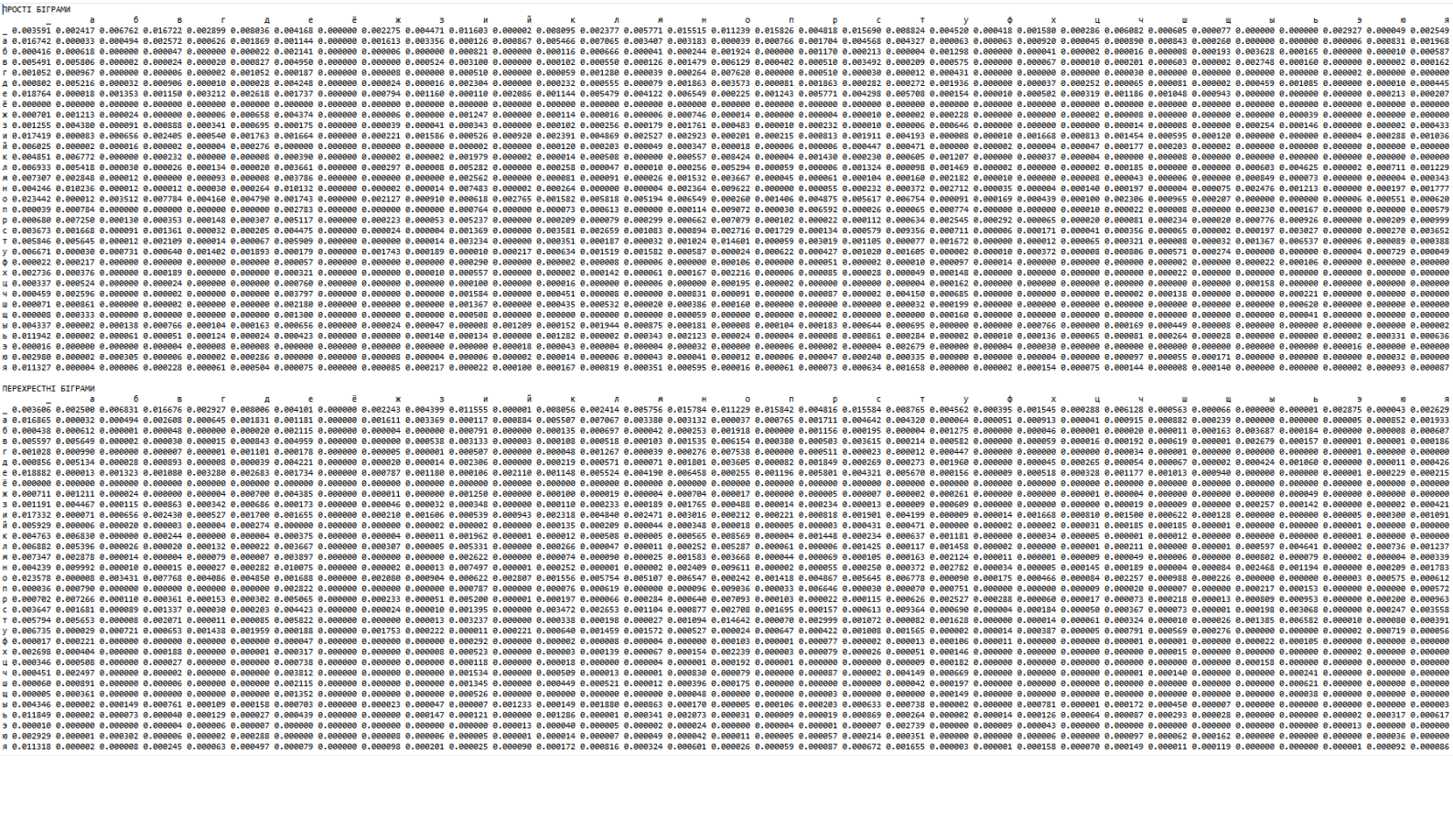
```
PS C:\Users\Данил\Desktop\CriptaLab1> python main.py
ЛІТЕРИ
ЕНТРОПІЯ: 4.449600333417484
о : 96681, е : 73380, а : 67116, н : 54835, и : 54644, т : 54553, с : 44600, в : 38984, л : 38734, р : 35246, к : 27833, д : 26983, м : 26495, у : 24993, п : 23121, ь : 19566, я
 : 18000, ч : 15249, б : 14655, г : 14232, ы : 13912, з : 12975, ж : 9615, й : 8437, х : 7171, ш : 6937, ю : 4733, э : 2971, щ : 2521, ц : 2336, ф : 1049, ё : 65,

ПРОСТІ БІГРАМИ
ЕНТРОПІЯ: 8.251331417953548

ПЕРЕХРЕСТНІ БІГРАМИ
ЕНТРОПІЯ: 8.254367201998619
PS C:\Users\Данил\Desktop\CriptaLab1>
```

Для матриць з біграмами було виділено два файли, test_gap.txt та test.txt. Туди я повиводив значення частот для біграм у формі матриць, де на перехресті літер стоять біграми.

Без пробілів.



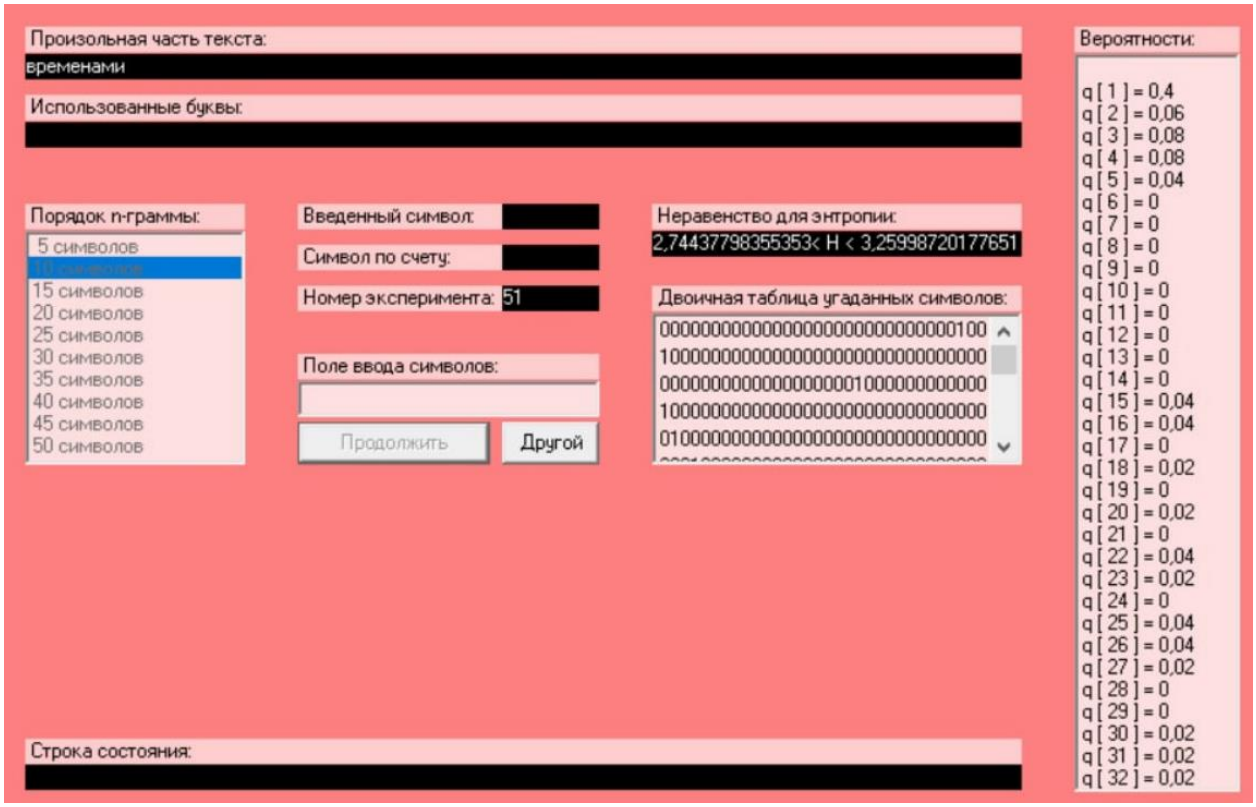
З пробілами.

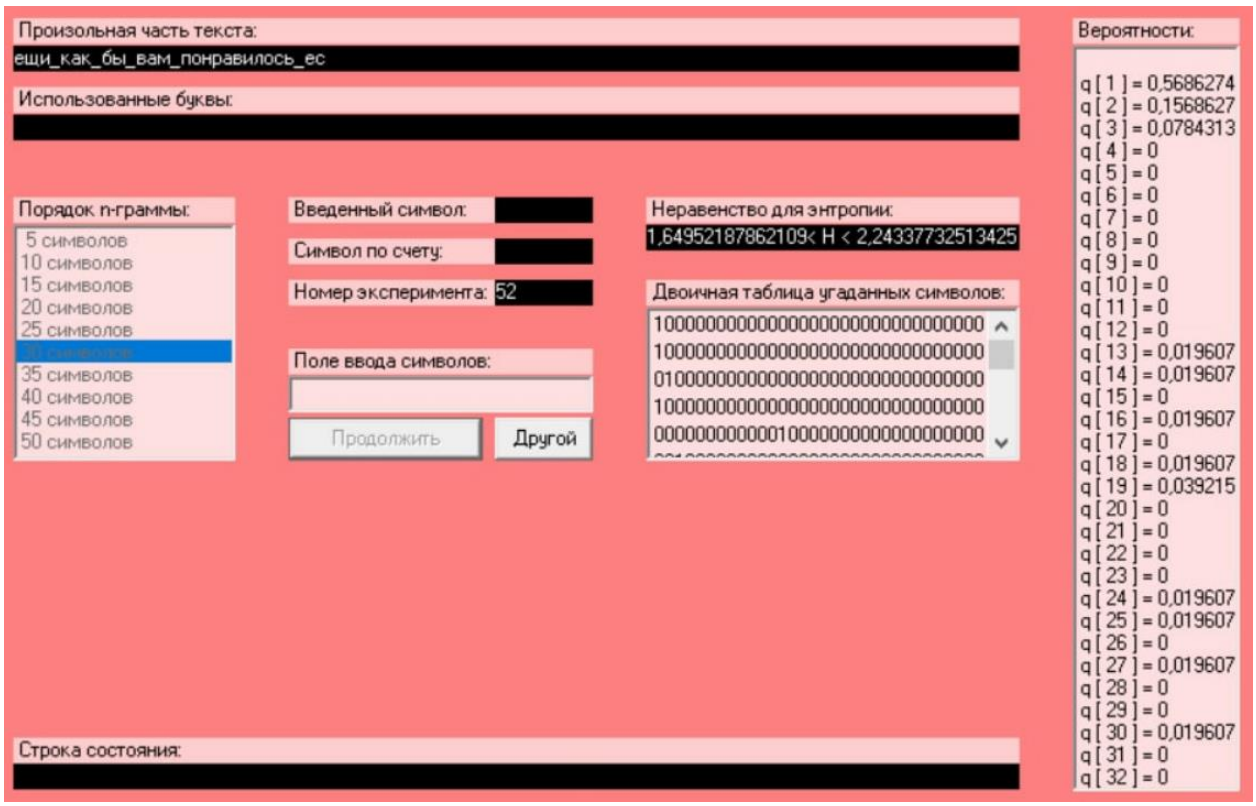


Ентропія	Значення ентропії	Надлишковість
Літери	4.3497	0.1300
Біграми	3.9496	0.6050
Перехресні біграми	3.9502	0.6049

II.

Для наступного завдання я використовую надану мені CoolPinkProgram. З допомогою цієї програми нам потрібно порахувати ентропію російської мови, а з нею і її надлишковість H(10), H(20), H(30).





H(10) = 3; H(20) = 2,8314; H(30) = 1,9464

Для обрахунку ентропії, я просумував усі значення ентропій і поділив їх на 3.

H = (3 + 2,8314 + 1.9464)\3 = 2,5926

Ентропія	Значення ентропії	Надлишковість
H(10)	3.0000	0.4000
H(20)	2.8140	0.4372
H(30)	1.9460	0.6108
H	2,5926	0.4800

Надлишковість рахується за формулою 1 – (H\H(0)), де H(0) для російської мови рівне 5. У мене це вийшло 0,48, я думаю такт мала надлишковість зумовлена моїми неякісними знаннями російської мови.

Висновок:

В ході роботи я навчився обраховувати ентропію джерела тексту для символів та біграм, також я навився обраховувати надлишковість тексту. Для російської мови надлишковість склала 48%, за вказаними вище причинами. Для пошуку ентропії біграм та літер було обрано “Злочин і кара”. Якщо не забрати пробіли то, тоді вони будуть найбільш вживаним символом (“_” : 172811), якщо ні то це літери: “о” : 96681, “е” : 73380 та “а” : 67116 відповідно, а най менш вживані: “ц” : 2336, “ф” : 1049. Найчастішими біграмами з пробілом виявились: “о_”: 0.023578, “е_”: 0.018882, “и_”: 0.017332, а без пробілів: “то”: 0.018084, “ов”: 0.012589, “не”: 0.012239. Наявність пробілів дуже полегшує роботу криптоаналітика.